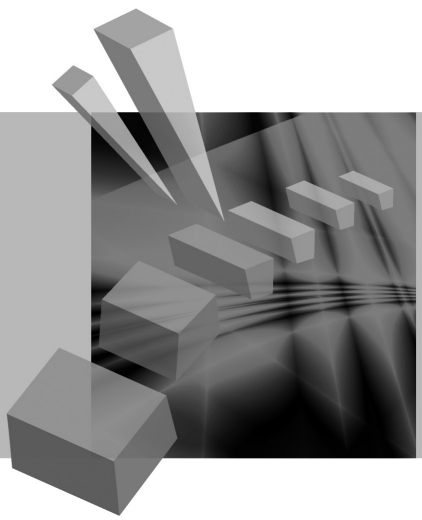
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Maciej Beręsewicz, Tomasz Klimanek
Uniwersytet Ekonomiczny w Poznaniu

WYKORZYSTANIE ESTYMACJI POŚREDNIEJ UWZGLĘDNIAJĄCEJ KORELACJĘ PRZESTRZENNĄ W BADANIACH CEN MIESZKAŃ

Streszczenie: Artykuł przedstawia propozycję wykorzystania metod estymacji pośredniej (w tym także metody, która uwzględnia korelację przestrzenną) do oszacowania pewnych charakterystyk rynku nieruchomości w województwie wielkopolskim. W konstrukcji odpowiednich estymatorów statystyki małych obszarów autorzy postanowili wykorzystać, oprócz modeli przekrojowych, także najnowsze podejścia w estymacji pośredniej wykorzystujące zależności przestrzenne. Ze względu na utrudniony dostęp do danych transakcyjnych badania oparte zostały na danych ofertowych zawierających informację o lokalizacji nieruchomości w przestrzeni geograficznej (dane zorientowane przestrzennie).

Słowa kluczowe: statystyka małych obszarów, autokorelacja przestrzenna, analiza rynku nieruchomości.

1. Wstęp

W ostatnich latach obserwuje się wzrost zainteresowania metodami estymacji pośredniej w Polsce. Wiele z dotychczasowych zastosowań dotyczyło problemów zwiększenia pokrycia informacyjnego dla potrzeb statystyki publicznej [Dehnel, Gołata 2006; Gołata 2004; Klimanek, Szymkowiak 2012; Kubacki 2008; Paradysz 2003]. Jednakże od samego początku stosowania metod statystyki małych obszarów podkreślano, że naturalnym odbiorcą wyników szacunków wydaje się obok statystyki publicznej i instytucji samorządowych także szeroko rozumiany biznes.

Należy także zwrócić uwagę na fakt, że większość zastosowań statystyki małych obszarów dotyczy estymacji charakterystyk rynku pracy, demografii i statystyki społecznej; stosunkowo niewielka jest liczba artykułów poświęconych zastosowaniom estymacji pośredniej w badaniach biznesowych.

Celem artykuł jest zastosowanie estymacji pośredniej do oszacowania przeciętnej ceny metra kwadratowego mieszkań na rynku nieruchomości mieszkaniowych w Poznaniu w sierpniu 2012 r.. Zastosowane podejście może zdaniem autorów, stanowić alternatywę w przypadku braku dostępu do danych transakcyjnych. Ponadto, przyjmując założenie o występowaniu wpływu lokalizacji w mieście na średnią cenę

mieszkania, autorzy podjęli próbę zastosowania w konstrukcji estymatora modelu wykorzystującego autokorelację przestrzenną. Tego rodzaju podejście zostało opisane w projekcie EURAREA, ale liczba zastosowań modeli wciąż jest niewystarczająca i wynika głównie z niedostatecznego wykorzystania istniejących informacji geoprzestrzennych – informacji opartych na współrzędnych geograficznych obiektów lub pewnych szczególnych charakterystykach związanych z tymi punktami.

2. Opis procedury badawczej

Rynek nieruchomości charakteryzuje się utrudnionym dostępem do danych charakteryzujących zawierane transakcje kupna-sprzedaży. W Internecie natomiast znajduje się spora liczba portali zajmujących się przedstawianiem ofert dotyczących m.in. sprzedaży mieszkań na rynku zarówno pierwotnym, jak i wtórnym. Stanowią one cenne źródło informacji na temat mieszkań oferowanych przez osoby prywatne oraz pośredników. Portale zawierają ceny ofertowe, co oznacza, że mogę się różnić od cen transakcyjnych, jednak m.in. Narodowy Bank Polski publikuje ceny ofertowe jako element cyklicznych raportów dotyczących rynku nieruchomości[1], a także tworzy bazę rynku nieruchomości (BaRn), w której uwzględnia zmiany cen zarówno transakcyjnych, jak i ofertowych. Portale, takie jak Gratka, Domy.pl (przy współpracy Open Finance) czy OtoDom, na podstawie cen ofertowych tworzą indeksy cen mieszkań oraz analizują sytuację na rynku nieruchomości[2].

W związku z ograniczonymi możliwościami uzyskania informacji na temat cen transakcyjnych oraz wykorzystywaniem cen ofertowych w analizach GUS i NBP postanowiono przeprowadzić analizę z wykorzystaniem cen oferowanych na portalach internetowych. W tym celu w programie R z wykorzystaniem pakietów XML oraz RCurl został napisany program (potocznie nazywany „pająkiem internetowym") umożliwiający automatyczne pobieranie informacji o ofertach mieszkań z rynku pierwotnego i wtórnego. Działanie programu opiera się na następujących krokach:

1. Wejdź na stronę wyników mieszkań dla Poznania. Ustal **i** = 1 oraz **n**, które oznacza ostatnią stronę.
   - 1.1 Pobierz ze strony wszystkie linki, które dotyczą ofert mieszkań (**j = 1 … m**).
   - 1.2 Dla każdego linku (**j = 1 ... m)** z punktu 1.1 wejdź na stronę i pobierz informacje o mieszkaniu.
   - 1.3 Jeżeli pobrano informacje o wszystkich mieszkaniach z 1.1, wróć do punktu 1.

[1] Raporty można znaleźć na stronie: http://www.nbp.pl/home.aspx?f=/publikacje/rynek_ nieruchomosci/index1.html.

[2] Opracowania i raporty można znaleźć na stronach internetowych www.gratka.pl, www.domy.pl czy www.otodom.pl.

2. Przejdź na kolejną stronę wyszukiwań (**i = i + 1)**. Jeżeli nie jest to ostatnia strona (**i ≠ n)**, wróć do punktu 1, w przeciwnym wypadku przejdź do punktu 3.

3. Zakończ działanie pętli.

Program działa do momentu, aż odwiedzi wszystkie strony zawierające wyniki wyszukiwań zadanych na początku (Poznań, rynek wtórny) oraz podstrony zawierające informacje o mieszkaniach. Oferty mieszkań były opisane szeregiem zmiennych, m.in. takich jak ceny mieszkania, ceny metra kwadratowego, liczba pokoi, powierzchnia, jak również informacja na temat położenia (w postaci współrzędnych geograficznych). Osoby zajmujące się publikowaniem informacji o danej nieruchomości umieszczały dodatkowe informacje, które nie występowały dla wszystkich ofert (np. typ budynku, rok budowy, stan).

Na potrzeby artykułu z portalu Domy.pl pobrano 14 229 ofert dotyczących mieszkań z wtórnego rynku nieruchomości mieszkaniowych w Poznaniu w sierpniu 2012 r., które po zastosowaniu omówionych w kolejnym rozdziale technik czyszczenia danych ograniczono do 9952 ofert.

Pierwszym etapem czyszczenia uzyskanego zbioru danych była eliminacja obserwacji powtarzających się. Deduplikacji dokonano, porównując ID oraz linki do ofert mieszkań. Następnie usunięto z analizy oferty mieszkań, które nie zawierały informacji o powierzchni, liczbie pokoi oraz położeniu.

Kolejnym etapem było wykorzystanie metody Least Trimmed Squares znajdującej się w poleceniu *PROC ROBUSTREG* pakietu SAS w celu detekcji wartości odstających. Metoda ta została zaproponowana przez Rousseeuwa [1984] i ma następującą postać:

$$\hat{\theta}_{\mathrm{LTS}} = \arg\min_{\theta} Q_{LTS}(\theta)\,, \tag{1}$$

gdzie:

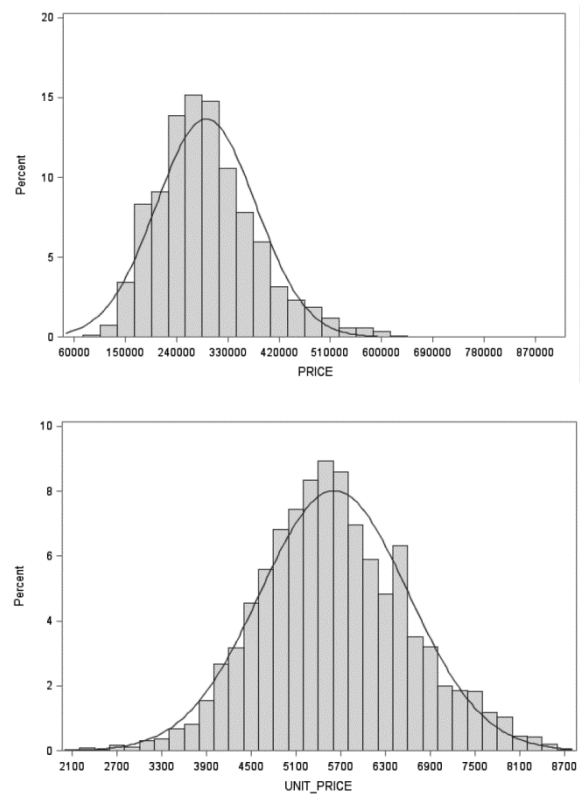
$$\mathrm{Q_{LTS}}(\theta) = \sum_{i=1}^{h} r_{(i)}^2, \tag{2}$$

$r_{(1)}^2 \le r_{(2)}^2 \le \cdots \le r_{(n)}^2$ są uszeregowanymi kwadratami reszt $r_i^2 = \left(y_i - x_i^T\theta\right)^2$, $i = 1, \ldots, n$, a h określone jest w przedziale $\frac{n}{2} + 1 \le h \le \frac{3n+p+1}{4}$.

W pakiecie SAS wartość progowa jest domyślnie ustawiona na $h = \frac{3n+p+1}{4}$.

Efektem etapu czyszczenia danych była baza, którą na potrzeby badania nazwano pseudopopulacją mieszkań w Poznaniu w sierpniu 2012 r. Na rysunku 1 zaprezentowano rozkład cen mieszkań oraz cen metra kwadratowego mieszkań w pseudopopulacji, tzn. po zastosowaniu procedur czyszczenia danych. Rozkład ceny metra kwadratowego jest bardziej zbliżony do rozkładu normalnego niż rozkład cen mieszkań.

W związku z faktem, że celem badania symulacyjnego miało być oszacowanie średniej ceny metra kwadratowego według jednostek ewidencyjnych miasta Pozna-

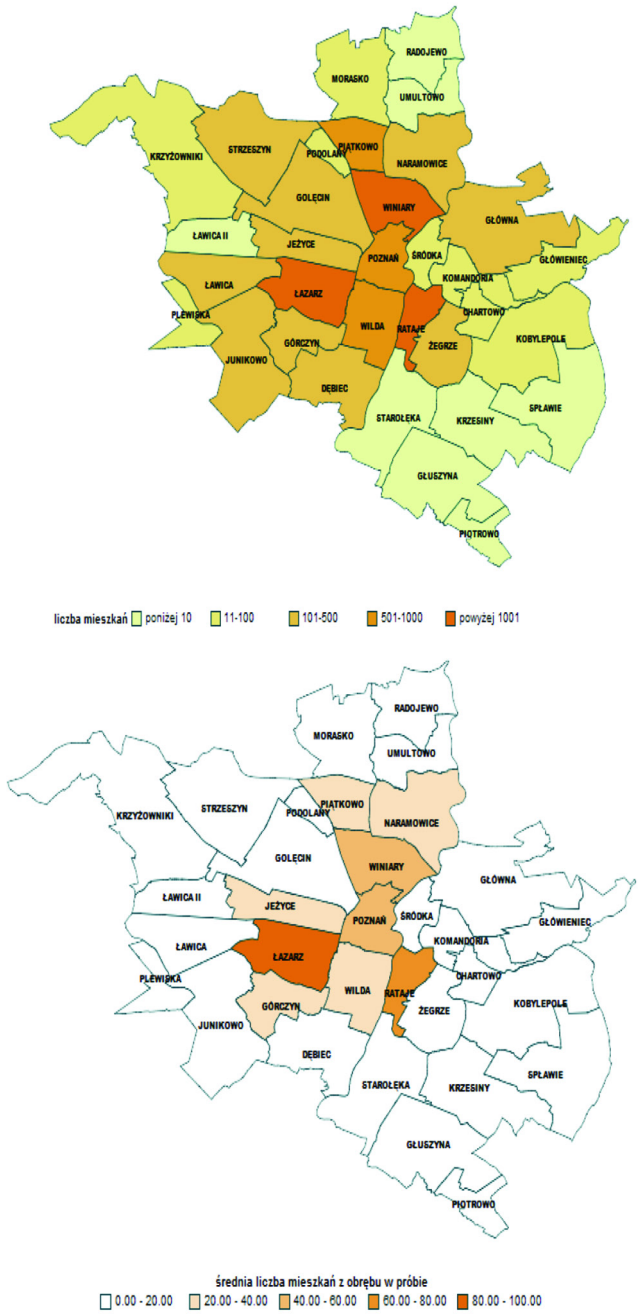


Rys. 1. Rozkład ceny oraz (PRICE) ceny m^2 (UNIT_PRICE) mieszkań w pseudopopulacji

Źródło: opracowanie własne w pakiecie SAS.

nia (obrębów[3]), a w populacji zawarta była jedynie informacja o współrzędnych geograficznych mieszkań znajdujących się na rynku ofert, należało przyporządkować mieszkaniom identyfikator obrębu. W tym celu posłużono się procedurą *PROC GINSIDE*, która sprawdza, czy podane współrzędne geograficzne mieszkania należą do określonego poligonu (obrębu miasta Poznania), a jeśli tak, to rekord w bazie zostaje uzupełniony o identyfikator tego poligonu. Następnie z populacji tej wylosowano 1000 prób o wielkości 5%, stosując schemat losowania prostego. Zastosowany schemat losowania spowodował, że w przypadku obrębów o niewielkiej liczebności mieszkań w populacji, w wylosowanych próbach reprezentacja mieszkań była niewielka lub nawet równa 0. Rozkład przestrzenny liczebności mieszkań w pseudopopulacji i badaniu symulacyjnym przedstawia rys. 2.

[3] Autorzy zdają sobie sprawę z faktu, że obręb geodezyjny nie jest najlepszym kryterium delimitacji przestrzeni dla wyznaczenia obszarów cenności, gdyż te przekraczają granice administracyjne (geodezyjne). Przyjęcie takiej jednostki przestrzennej wynikało z określonego zdefiniowania małego obszaru (domeny), dla której można było uzyskać podkład dla mapy numerycznej.



Rys. 2. Przestrzenny rozkład liczby mieszkań w pseudopopulacji i badaniu symulacyjnym

Źródło: opracowanie własne w pakiecie SAS.

Dla każdej z prób do oszacowania średniej ceny metra kwadratowego w obrębach miasta Poznania zastosowano 4 estymatory, przy czym w przypadku estymatorów typu GREG, EBLUP_B i SEBLUP, które wykorzystują zmienne pomocnicze z próby i spoza próby, do modelowania ceny metra kwadratowego wykorzystano powierzchnię mieszkania i liczbę pokoi. Należy zwrócić uwagę, że jakość modelu jest słaba, gdyż zmienność ceny metra kwadratowego została wyjaśniona przez wybrane zmienne objaśniające jedynie w około 12%.

Charakterystykę modelu w pseudopopulacji przedstawia tab. 1.

Tabela 1. Charakterystyka modelu w pseudopopulacji

Zmienna	Liczba stopni swobody	Oszacowanie parametru	Błąd standardowy	Statystyka *t*-Studenta	Prawdopodobieństwo testowe
Wyraz wolny	1	6745,65	32,58	207,05	<,0001
AREA	1	–16,00	0,98	–16,36	<,0001
ROOMS	1	–135,55	18,27	–7,42	<,0001

$R^2 = 0{,}123$, współczynnik zmienności losowej = 16,2%

Źródło: obliczenia własne w pakiecie SAS.

Zastosowano estymatory[4]:

- estymator bezpośredni (Horvitza-Thompsona)

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d}\sum_{i\in u_d} w_{id} y_{id} \,, \tag{3}$$

gdzie $\hat{N}_d = \sum_{i\in u_d} w_{id}$ oraz $w_{id} = \frac{1}{\pi_{id}}$

przy założeniu, że $\pi_{id,jd} = 0$ dla wszystkich $d \neq d'$ lub $i \neq j$;

- estymator GREG

$$y_{id} = \mathbf{x}_{id}^T \boldsymbol{\beta} + \varepsilon_{id} \,, \tag{4}$$

gdzie $E(\varepsilon_{id}) = 0, \quad Var(\varepsilon_{id}) = \sigma_\varepsilon^2$,

$$\hat{\bar{Y}}_d^{GREG} = \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{y_i}{\pi_i} + \left(\overline{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d}\sum_{i\in s_d}\frac{\mathbf{x}_i}{\pi_i}\right)^T \hat{\boldsymbol{\beta}} \,, \tag{5}$$

gdzie $\hat{N}_d = \sum_{i\in s_d}\frac{1}{\pi_i}$ i $\hat{\boldsymbol{\beta}}$ są oszacowane z wykorzystaniem ważonej metody najmniejszych kwadratów poprzez użycie wag wynikających ze schematu losowania:

[4] Wzory na oszacowania błędów średniokwadratowych zostały pominięte ze względu na ograniczania objętości tekstu niniejszej publikacji. Są one umieszczone w dokumentacji projektu EURAREA na stronie Urzędu Statystycznego Wielkiej Brytanii –http://www.statistics.gov.uk/eurarea.

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in u_d} w_{id} x_{id} x_{id}^T \right)^{-1} \sum_{i \in u_d} w_{id} x_{id} y_{id} \; ; \tag{6}$$

- estymator EBLUP_B będący kombinacją liniową estymatora bezpośredniego i syntetycznego (*EURAREA_Project_Reference_Volume* 2004),

$$\hat{\bar{Y}}_d^{EBLUP\_B} = \gamma_d \hat{\bar{Y}}_d^{DIRECT} + (1 - \gamma_d) \bar{X}_{.d}^T \hat{\beta} \tag{7}$$

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}, \text{ przy czym } u_d \sim iid\ N(0, \sigma_u^2), \quad e_{id} \sim iid\ N(0, \sigma_e^2)$$

$$\hat{\beta} = (x^T D^{-1} x)^{-1} x^T D^{-1} y,$$

gdzie: y – wektor obserwacji na zmiennej objaśnianej,

x – macierz o wierszach składających się z $\bar{x}_{.d}^T$,

D – macierz o iteracyjnie aktualizowanych elementach $(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$ na diagonali;

- estymator SEBLUP[5] uwzględniający autokorelację efektów losowych związanych z lokalizacją domen w przestrzeni [Saei, Chambers 2004; D'Alò, Falorsi, Solari 2004].
  W zapisie macierzowym model można zapisać następująco:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{8}$$

gdzie: $\boldsymbol{y}$ jest wektorem zmiennej objaśnianej, $\boldsymbol{X}$ i $\boldsymbol{Z}$ są znanymi macierzami rzędu odpowiednio: $\boldsymbol{N} \times \boldsymbol{P}$ (liczba obserwacji razy liczba zmiennych pomocniczych) i $\boldsymbol{N} \times \boldsymbol{D}$ (liczba obserwacji razy liczba małych obszarów). Macierz $\boldsymbol{Z}$ jest macierzą incydencji zdefiniowaną następująco:

$$\boldsymbol{Z} = \begin{bmatrix} 1_{N_1} & . & . & 0 \\ 0 & . & . & 0 \\ 0 & . & . & 0 \\ 0 & . & . & 1_{N_D} \end{bmatrix}, \text{ gdzie } 1_{N_d} \text{ jest wektorem o wymiarach } \boldsymbol{N_d}, \text{ którego}$$

wszystkie elementy są równe 1, $\boldsymbol{u}$ oraz $\boldsymbol{e}$ są wektorami zmiennych losowych o wartościach oczekiwanych równych 0 oraz macierzy wariancji – kowariancji odpowiednio:

$$\boldsymbol{N} \sim [0, \sigma_U^2 \boldsymbol{A}] \text{ oraz } \boldsymbol{N} \sim [0, \sigma^2 \boldsymbol{I}_N],$$

elementy $\boldsymbol{a}_{(dd')}$ macierzy $\boldsymbol{A}$ są dane wzorem:

[5] SEBLUP – Spatial EBLUP.

$$a_{(dd')} = \left[1 + \delta_{(dd')} \exp\left(\frac{dist(dd')}{\alpha}\right)\right]^{-1}, \tag{9}$$

gdzie: $dist(dd')$ oznacza odległość między małymi obszarami d i d'.

$$\delta_{(dd')} = \begin{cases} 0 & \text{for} \quad d = d' \\ 1 & \text{for} \quad d \neq d' \end{cases}, \tag{10}$$

a α jest parametrem skali.

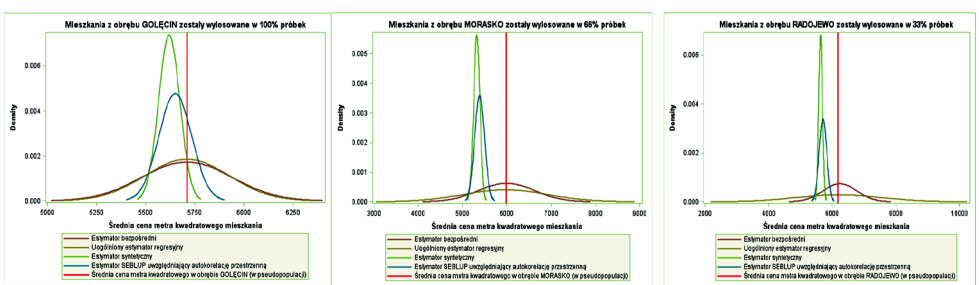
3. Uzyskane wyniki

Ze względu na ograniczania objętości tekstu niniejszej publikacji wyniki zostaną przedstawione w sposób bardzo syntetyczny[6]. Jednostki ewidencyjne (obręby) zostały podzielone na trzy kategorie:

Kategoria A – liczebność próby we wszystkich symulacjach przekraczała 3 mieszkania.

Kategoria B – liczebność próby w ponad 50% symulacji przekraczała 3. Kategoria C – liczebność próby w więcej niż 50% symulacji była mniejsza bądź równa 3.

Dla wymienionych powyżej kategorii wybrano po jednym przykładzie obrębu i dokonano oceny obciążenia 4 zastosowanych estymatorów i oceny wzrokowej wariancji tych estymatorów. Podkreślić jednak należy, że charakterystyki rozkładów są podobne dla pozostałych obrębów w ramach danej kategorii, stąd w ocenie autorów zaprezentowane przypadki są dobrymi reprezentantami obrębów w poszczególnych kategoriach.



Rys. 3. Rozkład estymatorów w obrębie Golęcina (kat. A), Moraska (kat. B) oraz Radojewa (kat. C)

Źródło: opracowanie własne w pakiecie SAS.

[6] Autorzy dysponują rozkładem estymatorów dla wszystkich 33 analizowanych obrębów, ale ograniczenia publikacji spowodowały to, że możliwa jest prezentacja jedynie wybranych jednostek ewidencyjnych.

4. Wnioski

Przeprowadzone badanie symulacyjne dostarczyło pewnych wniosków, które można sformułować następująco:

- estymator bezpośredni, chociaż nieobciążony, charakteryzuje się dwiema podstawowymi wadami w przypadku estymacji dla małych domen:
  1) ma nieakceptowalną wariancję, a w związku z tym także błąd szacunku,
  2) w przypadku zerowej próby w domenie nie można wyznaczyć oceny estymatora;
- uogólniony estymator regresyjny, chociaż umożliwia uzyskanie oceny estymatora w przypadku zerowych prób, to jednak charakteryzuje sie równie dużą wariancją co estymator bezpośredni;
- estymatory syntetyczne i uwzględniające autokorelację przestrzenną charakteryzują się niewielką wariancją. W porównaniu do estymatorów bezpośrednich są one jednak obciążone;
- analiza przestrzennego rozkładu estymatora uwzględniającego autokorelację przestrzenną sugeruje, że może być on dobrym narzędziem do szacowania cen nieruchomości w domenach, którymi są na przykład części miast o niewielkiej liczbie mieszkań na wtórnym rynku nieruchomości mieszkaniowych.

Literatura

AiM Property Małeccy Adamiczka spółka jawna, *Analiza cen transakcyjnych lokali mieszkalnych*, Poznań 2009.

D'Alò M., Falorsi S., Solari F., *EURAREA Documentation on SAS/IML program on Linear Mixed Model with Spatial Correlated Area Effects in Small Area Estimation*, EURAREA Deliverable, http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/downloads/index.html, 2004

Dehnel G., Gołata E., *Attempts to estimate basic information for small business in Poland*, "Statistics in Transition", Główny Urząd Statystyczny, Warszawa 2006, vol. 6, Number 5, s. 755-776

EURAREA_Project_Reference_Volume, http://www.statistics.gov.uk/eurarea, 2004.

Gołata E., *Problems of estimate unemployment for small domains in Poland*, "Statistics in Transition", Główny Urząd Statystyczny, Warszawa 2004, vol. 6, Number 5, s. 755-776.

Instytut Ekonomiczny Narodowego Banku Polskiego, *Raport o sytuacji na rynku nieruchomości mieszkaniowych i komercyjnych w Polsce w 2011 r.*, NBP, Warszawa 2012

Klimanek T., Szymkowiak M., *Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy,* [w:] Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 2012, s. 601-609.

Kubacki, J., *Application of Bayesian estimation methods for small domains in the Polish Labor Force Survey*, Acta Universitatis Lodziensis, Folia Oeconomica 216, Łódź 2008, s. 389-396

Paradysz J., *Zasilanie publicznej statystyki regionalnej za pomocą estymacji dla małych obszarów w perspektywie wykorzystania rejestrów administracyjnych*, „Wiadomości Statystyczne", Główny Urząd Statystyczny, Warszawa 2003, nr 4, s. 1-9.

Rao J.N.K., *Small Area Estimation*, John Wiley & Sons, Inc, 2004.

Rousseeuw P.J., *Least median of squares regression*, "Journal of the American Statistical Association" 1984, 79, 871-880.
Saei A., Chambers R., *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, University of Southampton, 2004.

USING INDIRECT ESTIMATION WITH SPATIAL AUTOCORRELATION IN DWELLING PRICE SURVEYS

Summary: The article presents the application of indirect estimation methods (including the method accounting for spatial correlation) to estimate some characteristics of real estate market in Wielkopolska Voivodeship. To build the small area estimators the authors decided to apply not only cross-sectional models but also the most up to date approach using spatial correlation. Because the access to the transactional data was not possible the research was based on the Internet data (offers) which included information about localization (spatially oriented data).

Keywords: small area statistics, spatial autocorrelation, real estate market analysis.